PROJECT REPORT: PREDICTING BUSINESS SUCCESS BASED ON YELP REVIEWS

Executive Summary

This project aims to predict business success, defined as achieving an average Yelp rating above 4 stars, using a dataset from Yelp. The study incorporates sentiment analysis, review length, user interactions, and business attributes to derive actionable insights. Big data technologies and machine learning were employed, with a focus on **Apache Spark** for data processing and analysis.

Key Objectives

1. Identify features in customer reviews (e.g., sentiment, length) that predict success.
2. Explore the impact of business attributes like review count.
3. Build a predictive model to classify businesses as successful or not based on ratings.

Data Source

- **Yelp Academic Dataset:**
  - Files: business.json, review.json, checkin.json, user.json, and tip.json.
  - **Storage and Processing:** Data stored in Hadoop HDFS and processed using Apache Spark.
- **Challenges:** Addressed missing values, duplicate records, and performance optimization during large-scale data processing.

Data Insights

1. **Star Ratings Distribution:**
   - The majority of businesses have ratings between 3 and 4 stars, with fewer at the extremes (1 and 5 stars).
2. **Business Categories:**
   - Restaurants dominate the dataset, followed by Food and Nightlife categories.
3. **Sentiment Trends:**
   - Positive sentiment is slightly higher in spring and winter, with sentiment scores clustering near 1.0.
4. **Open vs. Closed Businesses:**
   - 82.5% of businesses are closed, highlighting significant market churn.
5. **Review Count vs. Star Ratings:**
   - Higher review counts do not necessarily lead to better ratings.

Model Development

Feature Engineering:

- Features: Sentiment score, review length, user interactions (useful, funny, cool votes), review count, and attribute counts.

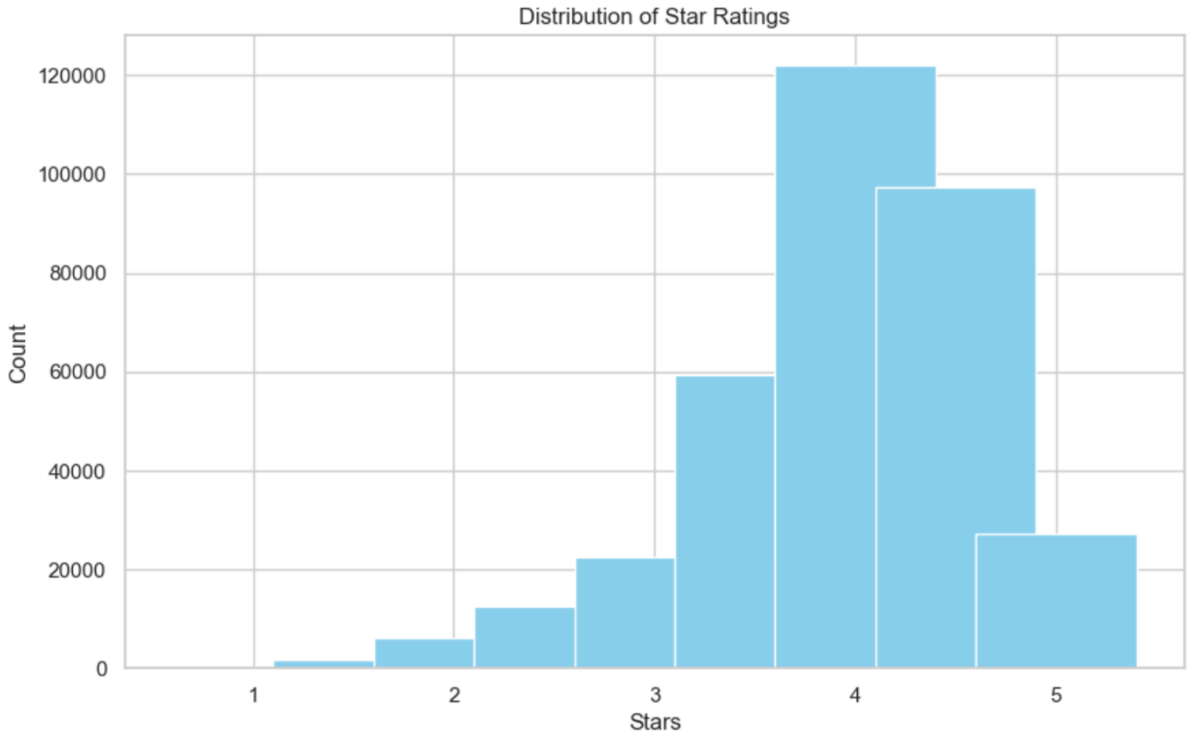
Machine Learning Pipeline:

1. **Data Preparation:**
   - Used **VectorAssembler** in PySpark to combine features into a single vector column.
2. **Model:**
   - Logistic Regression implemented using **PySpark MLlib**.
3. **Performance Metrics:**

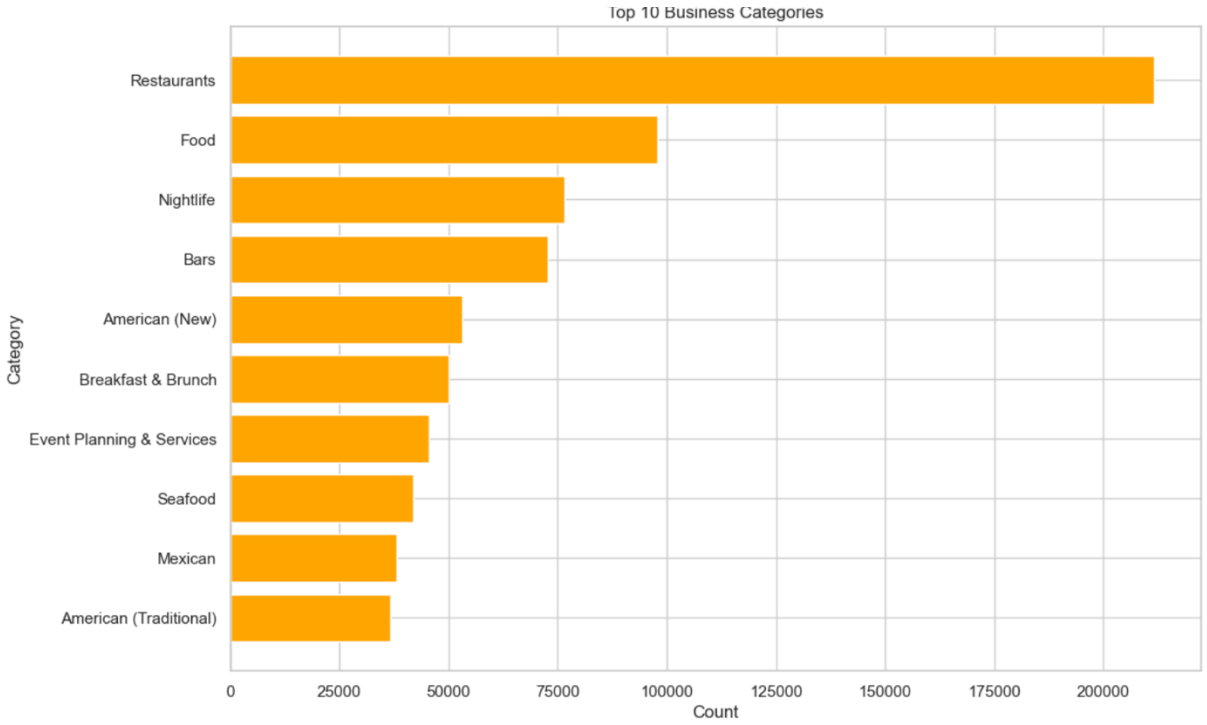
- **Accuracy:** 66.06%
- **F1 Score:** 63.01%
- **Model Coefficients:**
  - Sentiment Score: **0.997** (strong positive predictor).
  - Review Length: **-0.00026** (minimal negative impact).
  - User Interactions: Mixed correlations with varying strengths.

Visualizations

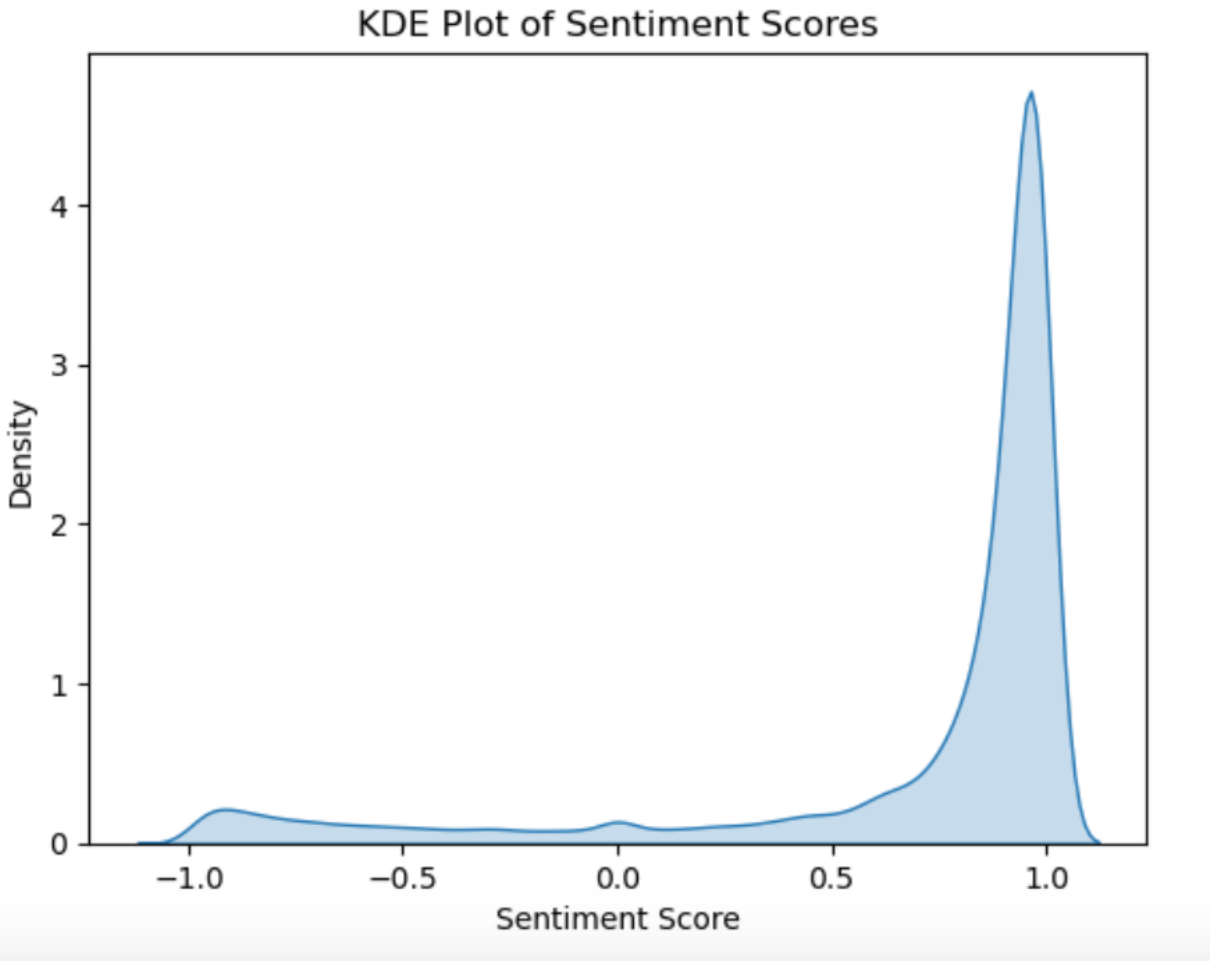
1. **Star Rating Distribution:**
   - The bar plot shows the distribution of star ratings, with most businesses clustered around 3 and 4 stars. Ratings below 2 and above 5 are relatively sparse.



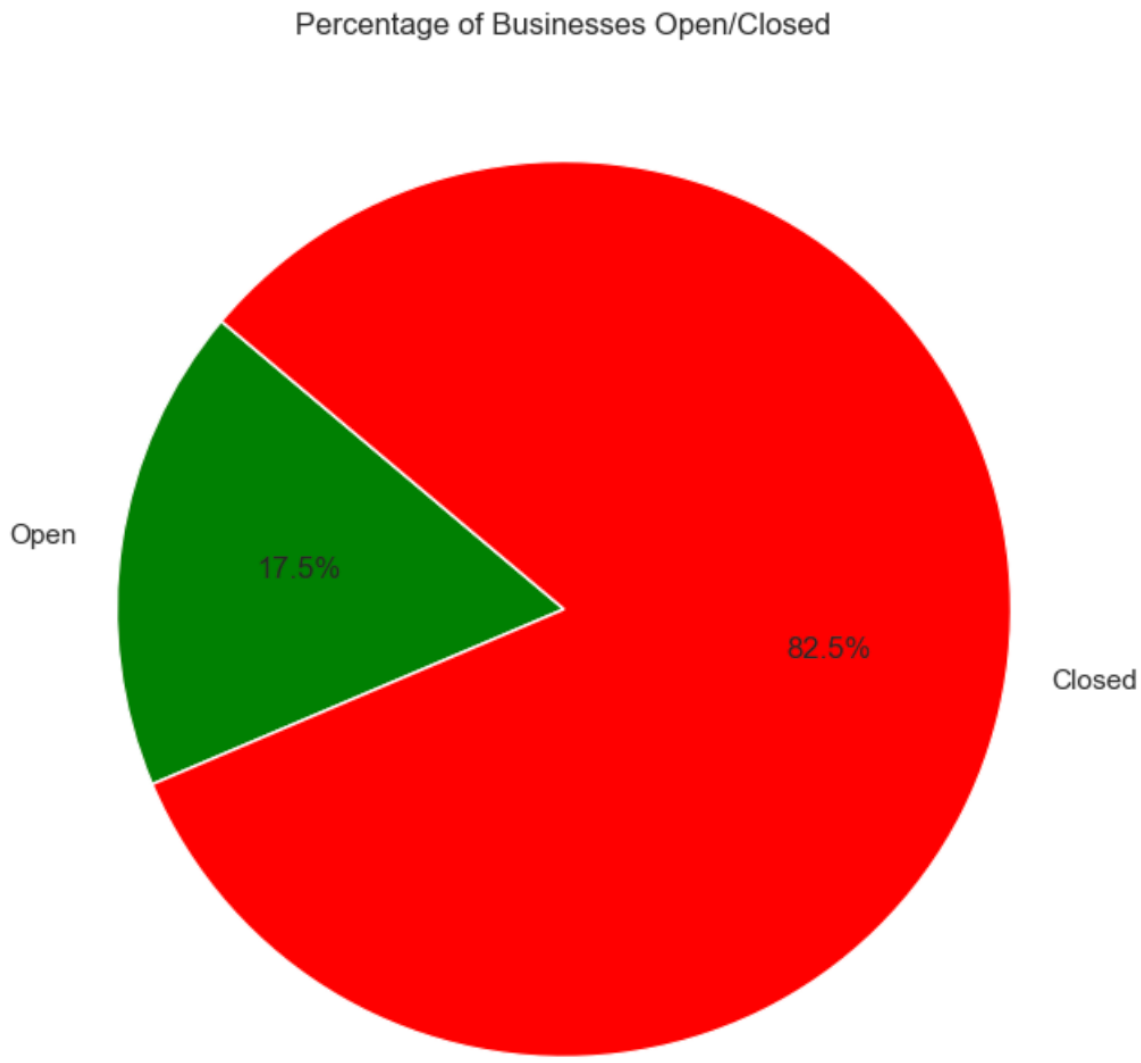
2. **Top Business Categories:**
   - A bar chart highlighting the most common business categories. Restaurants dominate, followed by Food, Nightlife, and Bars, reflecting the focus of Yelp users.



3. **KDE Plot of Sentiment Scores:**
   - A kernel density estimate (KDE) plot illustrating the distribution of sentiment scores.
   - Most reviews have highly positive sentiment (near 1.0), with very few reviews in the neutral or negative range (below 0).



4. **Open vs. Closed Businesses:**
   - A pie chart showing the percentage of open and closed businesses.
   - 82.5% of businesses in the dataset are closed, indicating significant churn in the market.



5. **Review Count vs. Star Ratings:**
   - A scatterplot showing the relationship between review count and star ratings.
   - While some high-review businesses have higher ratings, the relationship is scattered, indicating that review count alone does not predict star ratings.

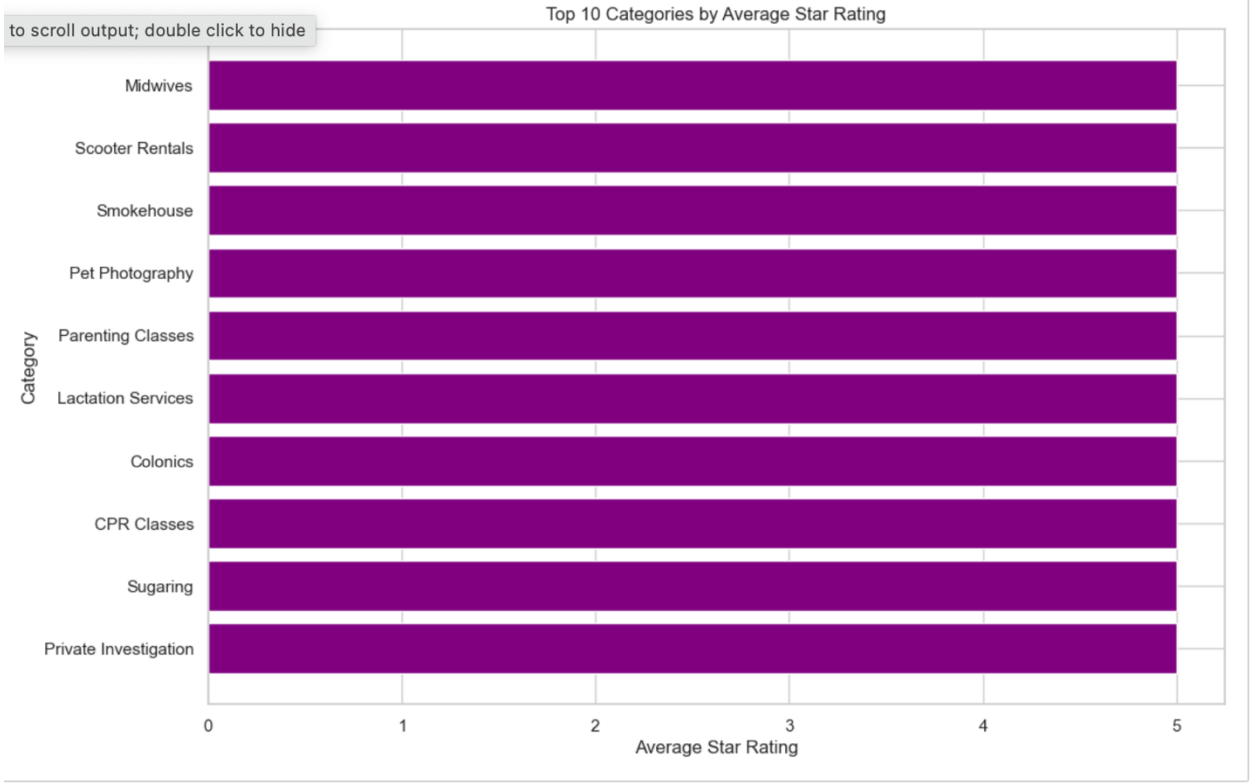
6. **Sentiment Trends (Monthly):**
   - Sentiment is slightly higher in spring and winter, with minor dips in summer and fall.

month	avg_sentiment
1	0.6898261686955284
2	0.6885068657765205
3	0.6879969909661376
4	0.6877867737124703
5	0.680016162772347
6	0.6766812795050746
7	0.6768130037255146
8	0.6785527398424621
9	0.6743692491798972
10	0.6765039576779165
11	0.6833621641632577
12	0.674251033252543

7. **Sentiment Trends (Seasonal):**
   - Spring and winter have the highest average sentiment scores, while summer has slightly lower sentiment.

```
+------+------------------+
|season|     avg_sentiment|
+------+------------------+
ck to scroll output; double click to hid
|Spring|0.6851282775497658|
|Summer| 0.677371333111911|
|Winter|0.6845174595000563|
+------+------------------+
```

8. **Top Categories by Average Star Rating:**
   - A horizontal bar chart showing the top 10 categories with the highest average star ratings.
   - Categories like Midwives, CPR Classes, Smokehouse, and Lactation Services have perfect or near-perfect average ratings.



Key Findings

1. **Sentiment Influence:** Sentiment score significantly impacts star ratings, emphasizing the importance of customer experience.
2. **Business Longevity:** The high closure rate suggests the need for strategic interventions to maintain operations.
3. **User Engagement:** Useful votes positively correlate with success, while funny and cool votes have minor impacts.

Limitations

- **Model Performance:** The logistic regression model achieved moderate accuracy, suggesting the need for more advanced models.

- **Feature Set:** Incorporating additional features like time of review and user demographics could improve predictive performance.

Future Work

1. Experiment with ensemble methods (e.g., Random Forest, Gradient Boosting).
2. Integrate time-series analysis to identify trends and seasonality in reviews.
3. Expand the feature set with attributes like user demographics and business hours.

Conclusion

This project demonstrates the utility of leveraging Yelp data and Apache Spark for scalable data analysis and machine learning. The findings provide valuable insights for business owners and a foundation for future research in predicting business success.